"Prediction of Sleepiness Ratings from Voice by Man and Machine": A Perceptual Experiment Replication Study

Vincent P. Martin, Aymeric Ferron, Jean-Luc Rouas, Pierre Philip

ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096193

"PREDICTION OF SLEEPINESS RATINGS FROM VOICE BY MAN AND MACHINE": A PERCEPTUAL EXPERIMENT REPLICATION STUDY

Vincent P. Martin[⋆†] *Aymeric Ferron*[⋆] *Jean-Luc Rouas*[⋆] *Pierre Philip*[†]

[⋆] LaBRI, Univ. Bordeaux, CNRS, Bordeaux INP, UMR 5800, F-33400 Talence, France
[†]SANPSY, Univ. Bordeaux, CNRS UMR 6033, F-33076 Bordeaux, France

ABSTRACT

Following the release of the SLEEP corpus during the Interspeech 2019 paralinguistic continuous sleepiness estimation challenge, a paper presented at Interspeech 2020 by Huckvale *et al.* examined the reasons for the poor performance of the models proposed for this task. Careful analyses of the corpus led to the conclusion that its bias makes it hazardous to use for training machine learning systems, but a perceptual experiment on a subset of this corpus seemed to indicate that human hearing is however able to estimate sleepiness on this corpus.

In this study, we present the results of the Endymion replication study, in which the same samples were rated by thirty French-speaking naive listeners. We then discuss the causes of the differences between the two studies and examine the effect of listener and sample characteristics on annotation performances.

Index Terms— Sleepiness, Voice, Perceptual study, Experimental study replication, Paralinguistic

1. INTRODUCTION

Sleepiness is a complex psychophysiological phenomenon that has negative consequences on both personal and public health and increases the risk of disability and mortality [1]. The significant imbalance between the number of sleep specialists and the prevalence of sleepiness (up to 40% of the general population [2]) and the need for physicians to better follow up their patients between consultations has led to the adoption of Ecological Momentary Assessment (EMA). Thanks to EMA, clinicians have access to patients' symptoms very regularly in the usual living conditions of the patient, paving the way to personalized treatments and realtime relapse prevention [3].

A promising tool for measuring sleepiness in EMA is the voice. Indeed, it is linked to the speaker's physiological state [4] and it is possible to implement voice measurements in passive situations without requiring the patient to perform a specific task (e.g. interaction with a connected device). In this way, sleepiness detection in speech has been at the heart of two international challenges proposed in parallel with the 2011 and 2019 Interspeech conferences.

During the Interspeech 2011 challenge on speaker's state estimation [5], the Sleepy Language Corpus has been introduced. On this corpus containing the recordings of 99 speakers, the winner of the challenge achieved an Unweighted Average Recall (UAR) of 71.7% [6]. A more recent work on this corpus achieved an UAR of 77.6% using acoustic features [7].

More recently, the SLEEP corpus was published for the INTERSPEECH 2019 Computational Paralinguistics Challenge for continuous sleepiness estimation [8] with the new challenge of estimating sleepiness (correlation between predictions and ground truth values). It contains 16,492 random samples from 915 German-speaking people whose sleepiness levels are annotated with the Karolinska Sleepiness Scale [9, KSS]. The ground truth of the SLEEP corpus is the truncated average of three KSS: one is filled in by the subjects themselves, while the other two are annotated by assistants using any information they have available, including audio [10]. For more information, this corpus is described in details in [11]. Contrary to the expectations of the challenge organizers, the proposed systems did not show much improvement from the baseline ($\varrho = 0.387$ for the best system [12] vs. $\varrho = 0.343$ for the baseline). Even more recent works on this corpus using the latest deep learning techniques did not perform better [13, 14].

To investigate the causes of this glass ceiling, Huckvale *et al.* [15] conducted a perceptual study to test the suitability of the corpus for the proposed regression task. Based on the annotations of 90 samples extracted from the SLEEP corpus by 26 British English listeners, and using Wisdom of the Crowd, they achieved performances far beyond those ever achieved by the systems proposed for sleepiness estimation tasks (r = 0.72). Thus, we claim that the study by Huckvale *et al.* supports the hypothesis that the human ear can estimate sleepiness from speech samples of the SLEEP corpus.

We propose in this article to reproduce the perceptual study conducted in [15] (denoted as "original study") with

This work was partially funded by the SOMVOICE project sponsored by the Labex BRAIN (ANR-10-LABX-43). We are deeply indebted to all the participants who spent time participating in the Endymion study. We gratefully thank Pr. Jarek Krajewski for having given us access to the SLEEP corpus, and Pr. Mark Huckvale for having given us the necessary material for the replication of his study.

naive French-speaking listeners to confirm or infirm this hypothesis. This paper is organized as follows. In Section 2, we introduce the methodology of our replication study on the SLEEP corpus. We present our results in Section 3 and discuss them in Section 4. Finally, we draw conclusions in Section 5.

2. METHOD

Thirty French-speaking listeners were recruited by word of mouth to rate 100 samples extracted from the SLEEP corpus. None of the participants had any hearing impairment, and their understanding of German and their musical sensitivity were collected. All information about listeners available in both the original and replication studies is presented in Table 1. Using a KSS annotation tool, they annotated the same 100 samples of the SLEEP corpus as in the original study. Ten samples (one from each ground truth level) were selected for training, the remaining 90 formed the testing subcorpus. The order of the samples is the same in both studies, and the samples are shown and annotated one at a time. The annotation tools used in each study are shown on Figure 1. While the version used in the original study combines at the same time a Lickert-like scale (gradual textual description) and an Visual Analog Scale (continuous line with two anchors), the scale used in our replication study is a real Lickert scale presented in the same manner as it has been to speakers.

Characteristic	**Huckvale *et al.*** $n = 26$	**Endymion** $n = 30$
Age	18-60	20-60
Sex	-	M: 17 F: 13
Impairments in hearing	None	None
Native language	English	French
German language level	German $\neq$ first language	"Not at all" ($n = 19$) "At least a little" ($n = 11$)
Specific Musical Sensibility	-	No ($n = 16$) Yes ($n = 14$)
Compensation	£5 ($n = 20$) attendance credits ($n = 6$)	None

Table 1. Listeners' characteristics

3. RESULTS

In this Section, we run the same analysis as in the original study. The comparison between the two studies is reported in Table 2.

Metric	**Huckvale *et al.***	**Endymion**
	Z-scaled annotations	
Correlation Kendall's coefficient	$r = 0.249$ $\tau = 0.117$	$\boldsymbol{r = 0.318}$ $\boldsymbol{\tau = 0.23}$
	WoC z-scaled annotations	
Correlation Friedman test UAR F1 SL/NSL	$\boldsymbol{r = 0.72}$ 1|2,3,4,5,6,7|8,9 $\mathbf{93.6\%}$ $\mathbf{0.87/0.96}$	$r = 0.41$ 1|2,3,4|5,6,7|8,9 69.6% 0.51/0.81
	Complementary results (no normalization)	
ICC2-10 Std/listener *mean (std)*	0.668 1.83 (0.38)	**0.975** 2.34 (0.21)

Table 2. Comparison metrics between the original study and our replication study. *WoC*: Wisdom of the Crowd, *SL*: Sleepy, *NSL*: Not Sleepy

3.1. Z-normalized raw scores

First, the annotations are z-scaled per listener to eliminate the individual characteristics of the listeners. The distribution of the resulting annotations in both studies is shown in Figure 2 (left). The raw z-scaled annotations in the Endymion replication study resulted in slightly better Person and Kendall correlations than in the original study, but these achievements are still insufficient to accept the hypothesis that human hearing is able to estimate sleepiness from speech samples extracted from the SLEEP corpus.

3.2. Wisdom of the crowd

In a second step, a Wisdom-of-the-Crowd (WoC) procedure is applied: for each sample, all z-scaled annotations are averaged, yielding an average predicted score. The resulting distributions are shown on Figure 2 (right). Compared to the original study, applying WoC to listener annotations of the Endymion replication study brings a smaller gain in the correlation between estimated values and ground truth.

In order to determine underlying groups in the annotations, a Friedman test and the corresponding post-hoc analysis are calculated with the Python package `Pingouin` v.0.4.0 [16]. In the original study, the annotations are grouped into three sleepiness levels based on a Friedman test: a 'sleepy' group (KSS > 7), a 'normal' group ($2 \leq$ KSS ≤ 8) and an 'aroused' group (KSS $= 1$). The same analysis applied to our replication study suggests a division into four groups ($W = 0.263, ddof = 8, p = 0.007$, pairwise Mann-Whitney): a 'sleepy' and an 'aroused' groups (resp. KSS > 7 and KSS $= 1$), and two 'slightly aroused' and 'slightly sleepy' subgroups (KSS $\in \{2, 3, 4\}$ and KSS $\in \{5, 6, 7\}$). The sleepiness subgroups for each study can be observed on Figure 2 (right).

To calculate binary classification performance, the ground

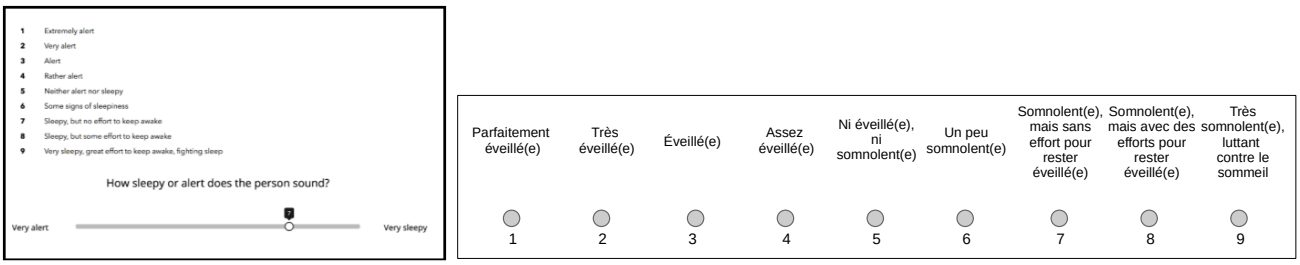

Fig. 1. KSS annotation tool proposed in the Huckvale *et al.* study (left, reproduced from [15]), and our replication study (right)

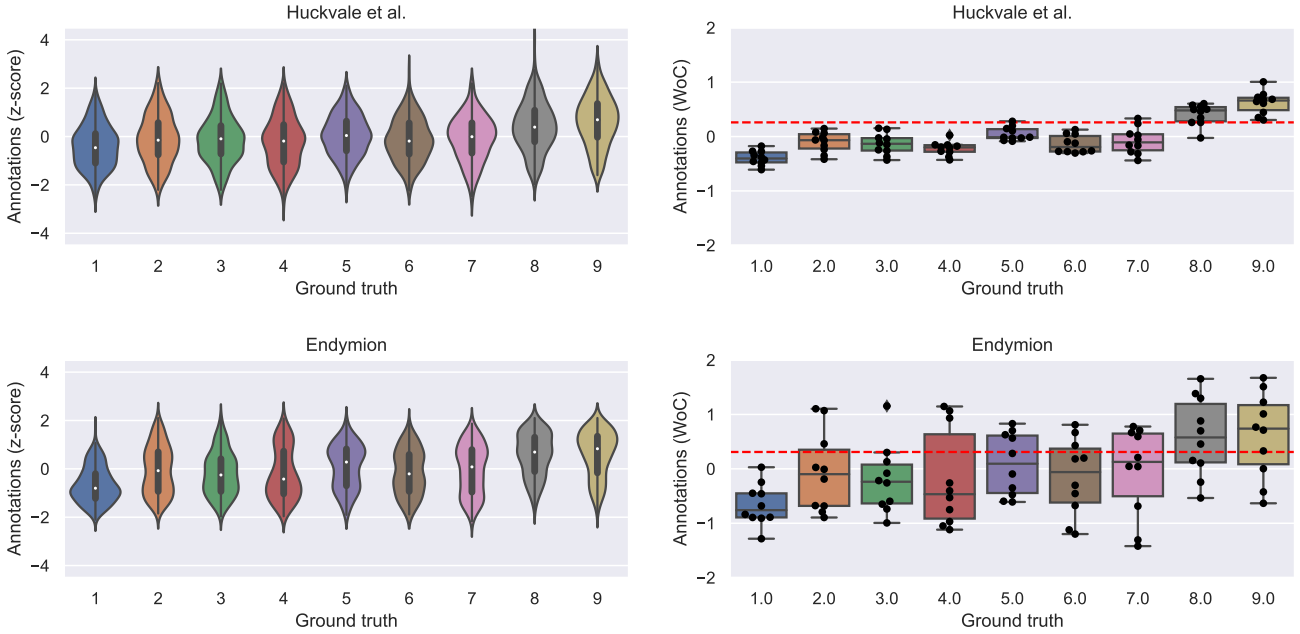

Fig. 2. (Left) Violin plots of z-scaled annotations by ground-truth value, in the study of Huckvale *et al.* (top) and in the Endymion study (bottom). (Right) Box plot of the WoC z-scaled annotations depending on the ground-truth KSS. Each dot represents a sample, and the red dashed line represents the cut-off value giving the best UAR

truth KSS is binarized into two classes: Sleepy (KSS $>$ 7) and Not Sleepy (KSS $\leq$ 7). A threshold of 0.26 on the z-scaled WoC annotation gives a binary classification UAR (Unweighted Average Recall) of 93.6% in Huckvale *et al.*, while in our replication study the best cut-off value of 0.31 yielded a UAR of only 69.6%, which is lower than classification performances typically achieved on the task. To complement these results for further discussion, we also calculated the F1 value for each class: in both studies, the F1 values are better in the NSL class than in the SL class.

4. DISCUSSION

4.1. Differences between the studies

To find the underlying cause of the differences between these analyses, we calculate the intra-class correlation (ICC2-10) on the raw annotations (before z-normalization) for each study. This metric assesses the reliability of the average ratings made by listeners and is an indicator of overall agreement between them [17]. It shows that there is a lower inter-annotator agreement in the original study (ICC $=$ 0.668) than in our replication study (ICC $=$ 0.975). This could be the source of the small performance gain bought by WoC in the Endymion study: averaging already converging opinions over a sample yields much less information than averaging dissenting opinions. To account for the variety of levels used by commenters in each study, we also calculates the standard deviation per listeners of their annotations (before z-normalization). The histograms of this metric in each study are shown on Figure 3.

The listeners of the original study use significantly fewer different levels than those of the Endymion study. However,

in the present study, listeners use a greater variety of levels, creating greater finesse in the annotations, albeit with less contrast (4 subgroups in the Endymion study vs. 3 subgroups in Huckvale *et al.*). These different behaviors may find their source in the presentation of the annotation scales between the two studies: while in the present study the textual description is directly above the selected value, some listeners may have inadvertently used the annotation scale in Huckvale *et al.* as a simple visual analog scale, without referring to the text description at the top of the screen, creating their own rating scale [18].

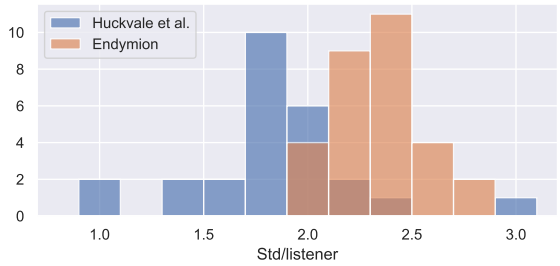

Fig. 3. Distribution of standard deviation of annotations per listener (before z-normalization). The observed difference is statistically significant (t-test, $p < 0.0001$)

4.2. What is the influence of the listeners' characteristics on their annotation performance?

In the Endymion study, two listener characteristics of particular interest for the task were collected: (1) their musical sensitivity, since music practice or a music-related hobby might improve the perception of speech [19, 20]; (2) their understanding of the German language, as the annotators who understand the language might have access to additional linguistic information. Therefore, for each annotator, we calculate the Mean Absolute Error (MAE) between the annotations she/he proposes (after z-scaling) and the corresponding ground truth values. Then, we min-max normalize them and we compute Mann-Whitney tests aiming to differentiate the MAE between each group.

We find no significant difference for either of the two previous factors (resp. $p = 0.190$ and $p = 0.228$ for musical sensitivity and German language comprehension). Thus, if some listener characteristics influence the way they estimate sleepiness from speech samples, they are not captured by our study.

4.3. Which samples are the hardest to annotate?

For a given sample, we hypothesize that two effects could explain the differences between annotations and ground truth. First, listener fatigue, who may not annotate the last samples as carefully as the first ones, for whom concentration may be easier (influence of the order of the samples). Second, the speaker's level of sleepiness, since the human ear might be able to detect some levels of sleepiness more easily than others (influence of the KSS). For each study and each sample, we computed the MAE between the ground truth values and the listeners' z-score annotation (MAE per sample). To make the ground truth and z-scale annotation values comparable, we min-max normalize them so that their minimum value is 0 and their maximum value is 1. Then we calculated the correlation (Spearman's ρ) between the MAE per sample and their order and between the MAE per sample and the ground truth KSS.

Factor	Huckvale *et al.*	Endymion
Order	$\varrho = 0.09, p = 0.39$	$\varrho = -0.13, p = 0.24$
KSS	$\varrho = 0.20, p = 0.05$	$\varrho = 0.40, p < 10^{-3}$

Table 3. Correlation between the MAE per sample and their order and ground truth KSS

The results are presented in Table 3. The MAE does not correlate with the sample index in either study, which excludes the hypothesis of listener fatigue. However, the per-sample MAE correlates weakly with the KSS ground truth of the original study and more strongly in the Endymion study: the higher the KSS, the larger the errors between annotations and ground truth. We put forward two hypotheses about this result. First, the human auditory system may be more sensitive to voice expressions of alertness than to sleepiness, explaining the increase in MAE with KSS. Coming back to the classification results from Section 3, the F1 values are better in the NSL class than in the SL class, which supports this hypothesis. Second, we cannot exclude the hypothesis that some speakers could have filled a KSS indicating a high sleepiness level at the time of their evaluation but were then stimulated by the various recording tasks. Therefore, sleepy subjects may make (involuntary) efforts to compensate for their sleepiness in order to complete the task, creating a difference between their self-reported level of sleepiness, assessed before the task, and the expression of their sleepiness in their voice.

5. CONCLUSION

To conclude, our replication study did not provide results as conclusive as the previous study conducted by Huckvale *et al.* [15]. Therefore, we cast doubt on the assumption that the human ear is capable of correctly assessing sleepiness from speech samples extracted from the SLEEP corpus. Regarding the factors influencing the annotation of the samples, we did not identify any influence of the annotators' characteristics. On the other hand, we found a link between the level of sleepiness of the speakers and the quality of sleepiness annotation in these two studies, with listeners having more difficulty in estimating very sleepy speakers.

6. REFERENCES

[1] Alexander J. Scott, Thomas L. Webb, Marrissa Martyn-St James, Georgina Rowse, and Scott Weich, "Improving sleep quality leads to better mental health: A meta-analysis of randomised controlled trials," *Sleep Medicine Reviews*, vol. 60, pp. 101556, 2021.

[2] Terry B. Young, "Epidemiology of daytime sleepiness: Definitions, symptomatology, and prevalence," *The Journal of Clinical Psychiatry*, vol. 65 Suppl 16, pp. 12–16, 2004.

[3] Saul Shiffman, Arthur A. Stone, and Michael R. Hufford, "Ecological Momentary Assessment," *Annual Review of Clinical Psychology*, vol. 4, no. 1, pp. 1–32, 2008.

[4] Guy Fagherazzi, Aurélie Fischer, Muhannad Ismael, and Vladimir Despotovic, "Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice," *Digital Biomarkers*, pp. 78–88, Apr. 2021.

[5] Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, and Jarek Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Interspeech 2011*, 2011, pp. 3201–3204.

[6] Dong-Yan Huang, Yu Tsao, Hori Chiori, and Hideki Kashioka, "Feature Normalization and Selection for Robust Speaker State Recognition," in *IEEE - International Conference on Speech Database and Assessments*, 2011.

[7] Vincent P. Martin, Jean-Luc Rouas, Pierre Thivel, and Jarek Krajewski, "Sleepiness detection on read speech using simple features," in *10th Conference on Speech Technology and Human-Computer Dialogue*, Timisoara, Romania, 2019.

[8] Björn Schuller, Anton Batliner, Christian Bergler, Florian B. Pokorny, Jarek Krajewski, Margaret Cychocz, Ralf Vollman, Sonja-Dana Roelen, Sebastian Schnieder, Elika Bergelson, Alejandrina Cristia, Amanda Seidl, Anne Warlaumont, Lisa Yankowitz, Elmar Nöth, Shahin Amiriparian, Simone Hantke, and Maximilian Schmitt, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," in *Interspeech 2019*, 2019.

[9] Torbjorn Åkerstedt and Mats Gillberg, "Subjective and objective sleepiness in the active individual.," *Int J Neurosci*, vol. 52, pp. 29–37, 1990.

[10] Florian Hönig, Anton Batliner, Elmar Nöth, Sebastian Schnieder, and Jarek Krajewski, "Acoustic-Prosodic Characteristics of Sleepy Speech - Between Performance and Interpretation," in *Speech Prosody 2014*, 2014, pp. 864–868.

[11] Vincent P. Martin, Jean-Luc Rouas, Jean-Arthur Micoulaud-Franchi, Pierre Philip, and Jarek Krajewski, "How to Design a Relevant Corpus for Sleepiness Detection Through Voice?," *Frontiers in Digital Health*, vol. 3, pp. 124, 2021.

[12] Gábor Gosztolya, "Using Fisher Vector and Bag-of-Audio-Words Representations to Identify Styrian Dialects, Sleepiness, Baby & Orca Sounds," in *Interspeech 2019*, 2019, pp. 2413–2417.

[13] Jose Vicente Egas-Lopez and Gabor Gosztolya, "Deep Neural Network Embeddings for the Estimation of the Degree of Sleepiness," in *ICASSP 2021*, Toronto, ON, Canada, 2021, pp. 7288–7292.

[14] Shahin Amiriparian, Pawel Winokurow, Vincent Karas, Sandra Ottl, Maurice Gerczuk, and Björn W. Schuller, "A Novel Fusion of Attention and Sequence to Sequence Autoencoders to Predict Sleepiness From Speech," arXiv 2005.08722, 2020.

[15] Mark Huckvale, Andras Beke, and Mirei Ikushima, "Prediction of Sleepiness Ratings from Voice by Man and Machine," in *Interspeech 2020*, 2020.

[16] Raphael Vallat, "Pingouin: Statistics in Python," *Journal of Open Source Software*, vol. 3, no. 31, pp. 1026, 2018.

[17] Terry K. Koo and Mae Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, June 2016.

[18] Carolyn C Preston and Andrew M Colman, "Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences," *Acta Psychologica*, vol. 104, no. 1, pp. 1–15, Mar. 2000.

[19] Salomi S. Asaridou and James M. McQueen, "Speech and music shape the listening brain: Evidence for shared domain-general mechanisms," *Frontiers in Psychology*, vol. 4, 2013.

[20] William Forde Thompson, E. Glenn Schellenberg, and Gabriela Husain, "Decoding speech prosody: Do music lessons help?," *Emotion*, vol. 4, no. 1, pp. 46–64, 2004.